Supplementary Appendix

Supplement to: Klauer SG, Guo F, Simons-Morton BG, et al. Distracted driving and risk of road crashes among novice and experienced drivers. N Engl J Med 2014;370:54-9. DOI: 10.1056/NEJMsa1204142

Distracted Driving Risk Among Novice and Experienced Drivers: Appendix

Sheila G. Klauer, Virginia Tech Transportation Institute, USA*

Feng Guo, Virginia Tech Transportation Institute, USA

Bruce G. Simons-Morton, National Institutes of Child Health and Human Development, USA

Marie Claude Ouimet, University of Sherbrook, Montreal, USA

Suzie E. Lee, Virginia Tech Transportation Institute, USA

Tom A. Dingus, Virginia Tech Transportation Institute, USA

Contents

Table of Figures

Table of Tables

Appendix 1: Naturalistic Driving Data Coding

Recent technological improvements have enabled traffic safety researchers to study driver behavior *in situ* or in real world traffic environments. Improvements in computer processing speed and data storage coupled with the reduction in physical size of these components have not only allowed instrumented vehicle studies to efficiently gather more parametric data, but have also allowed for vast improvements in video data collection. These improvements permit safety professionals to retrofit vehicles with state of the art eye tracking systems, physiological monitoring equipment, and collision warning systems, as well as facilitating the large-scale collection of driving performance data for long periods of time (e.g., one hundred vehicles for one year).

It is these recent technological improvements that will help to fill the lack of data regarding driver behavior in the seconds leading up to crashes and near-crashes. Thus safety professionals have been able to assess data on actual crashes (like epidemiological databases) with high resolution, detailed driving performance data (like empirical studies) as well as video data that records driver behavior as well as surrounding traffic. These instrumented vehicle studies are an important tool for researchers to add to their tool box for improving safety on our roadways.

This appendix will describe how the continuous video data can be efficiently reviewed and important data elements accurately recorded by trained coders. While many instrumented vehicle studies have been conducted with varying numbers of vehicles and lengths of data collection, this chapter will focus specifically on two larger-scale naturalistic driving studies: the 100 Car Naturalistic Driving Study (100 Car Study) and the Naturalistic Teenage Driving Study (NTD Study).

Data coding for both the 100 Car Study (Dingus, Klauer, Neale, Petersen, Lee, Perez, et al, 2006) and the NTD Study (Lee, Simons-Morton, Klauer, Ouimet, & Dingus, 2011) involved the coding of driver behavior prior to and during both crashes/near-crashes (CNC) but also during control road segments or normal, uneventful driving scenarios. For every CNC that occurred, two to three highly trained coders reviewed and recorded driver behavior. For the control road segments, multiple data coders reviewed the video and driving performance data and then recorded observed driver behaviors.

To identify the CNC events in these large datasets, triggering was performed using *a priori* determined threshold values for specific kinematic data variables. Triggering using the vehicle kinematic data identifies moments in time when the driver exceeds a set value (i.e., -0.55g longitudinal deceleration) or a combination of values (i.e., -0.55g longitudinal deceleration AND time-to-collision value of 2.0 seconds or less). Both researchers and practitioners have used kinematic triggered data to identify safety critical events such as crashes, near-crashes, or critical incidents or to identify ‘coachable’ events where drivers are exceeding the safe limits of their vehicle.

Using kinematic data to identify safety-relevant events is an intensive process. First, trained data reductionists view a sample of triggers created under a given set of criteria, and place each trigger into a “valid” or “invalid” category. For example, to create triggers to identify crashes and near-crashes, valid triggers represent any conflict that did or could have resulted in a crash. Invalid triggers may appear as normal driving and are often driver-specific (e.g., a driver that normally brakes hard) or as anomalies in the performance data. When criteria are tightened up to reduce the number of invalid triggers, this review of a sample of triggers is repeated. A cross check is also performed to ensure that the valid triggers found previously are still found with the more restrictive criteria and that the number of invalid triggers are reduced. When the criteria are finalized and the final triggers are generated, event classification begins.

Random selection of events can also be conducted to obtain a measure of ‘normal’ or baseline driving. This procedure was performed to identify the control road segments, as used in the referenced paper. The samples identified for the 100 Car Study were stratified based upon the number of hours traveled. The samples identified for the NTD Study were stratified based upon vehicle miles traveled, both of which are measures to ensure a normalized distribution of baseline samples.

After the sample of events were selected and identified, trained data coders then reviewed the events. Trained data coders first watched the appropriate video and the corresponding kinematic data for each of these events and verified that the event met all of the selection criteria and secondly, recorded the relevant event, driver behaviors, and environmental and roadway variables. These variables include:

- *Event Variables:* Variables used to establish the scenario and sequence of events prior to and through the CNC. These variables include event severity, event nature (e.g., conflict with lead vs. crossing vehicle), pre-incident maneuver, precipitating event, driver reaction, post-maneuver control, information about other drivers/vehicles/objects involved (e.g., type, position, maneuvers, impairments), and fault assignment.
- *Driver Variables:* Variables used to systematically describe driver state prior to and during the critical event. These include driver ID, driver behavior (e.g., speeding, aggressive driving), driver impairments (e.g., drowsiness, anger, substance abuse), secondary task engagement and duration (e.g., cell phone use), placement of hands on the wheel, visual obstructions, and seatbelt use.
- *Environmental Variables:* Variables used to describe environmental and/or roadway conditions consistent with GES and other crash databases. These include roadway surface condition, traffic flow, number of travel lanes, traffic density, traffic control device at event onset, relation to junction of vehicle at

event onset, roadway alignment (e.g., curve, grade), locality type (e.g., residential, interstate), ambient lighting, weather, and windshield wiper status.

Coder Training and Quality Control Policies

Figure 1 shows the Data Coding QA/QC Workflow that has been developed, tested, and implemented by the Virginia Tech Transportation Institute (VTTI). This workflow was used for the coding of both the 100 Car Study and for the NTD Study. The data coding QA/QC workflow has four phases, all of which are equally critical to the quality of manually coding data. These phases include Protocol Development, Coder Training, Data Coding, and Post-Coding with tasks assigned at each level of one of four roles. Each role and phase is discussed below and also illustrated in the diagram.

<<Insert Figure S1 about here>>

Roles:

Four roles are critical to the data coding process.

1. The Researcher or Research Manager (either internal or external) oversees the research project from research design through data collection, coding, analysis, and reporting. The researcher takes the lead in protocol and data dictionary development at the data coding phase, and provides input and feedback throughout all four phases.
2. The Data Coding Manager serves as the direct liaison between the researcher and the data coding team and oversees all QA/QC steps. Most questions from coders can be fielded by the Data Coding Manager; those that cannot are taken to the Researcher.
3. Senior Data Coders (or lab proctors) are generally experienced, highly trained data coders who monitor a project's progression through the QA/QC workflow, assist the Data Coder Manager with coder training, test new protocols before coding work begins, create and score tests to formally measure coder reliability, and monitor the workflow to coders.
4. Data coders perform the bulk of the data coding. They also participate in the QA/QC process by completing required tests and assisting with spot checks. Data coders are ideally limited to working no more than 4-5 hours per day, with a 10 minute break for every hour of work.

Phase 1: Protocol Development

Once the Researcher has drafted a preliminary protocol and data dictionary based on the research questions, it usually goes through several rounds of review with the Data Coder Manager. Due to the nature of the Data Coder Manager's work, this person often has more experience in finding potential ambiguities, knowing when categories may be missing from certain variables, and adapting new protocols to be consistent (if possible) with previous protocols for later cross-analysis. Once both the Researcher and the Data Coder Manager are satisfied with the draft protocol, it enters the first testing iteration. The Senior Coder takes the lead in this testing by viewing a variety of events and completing the coding for those events based on the draft protocol. As the Senior Coder works, he/she takes notes regarding uncertainties, what types of events were difficult to categorize, other variables that may seem important to the coding, and areas where the written protocol may need to be elaborated further. These comments are then reviewed by the Data Coder Manager and the Researcher.

Discussions at this point addressed how well the protocol performed in answering the research question, whether the data has been coded as intended, and whether this coding will provide the required information. If changes are significant, a second round of testing is recommended. Once the protocol is satisfactory, it enters the second phase, Coder Training.

As with any research, unexpected scenarios, driver behaviors, or environmental conditions often arise during data coding and/or protocol development, and this sometimes results in a need to edit, append, or further clarify the working data dictionary. For instance, if a road type, secondary task, or conflict type is observed in the video that does not clearly fit into the categories provided by the dictionary, then the data reduction manager, in conjunction with the researchers, must decide how to code that scenario for immediate and future reference. When this occurs, it is imperative that the previously established formal process for updating the data dictionary and/or data reduction manual be followed:

1. *Assess the situation.* Can the situation be described adequately by existing operational definitions given the research questions? (If no, continue to Step 2.)

2. *Consider the frequency.* How frequently will this situation arise? How critical is that situation to the outcome of the event? Does it warrant a new category and definition, or can existing operational definitions be modified to include this new situation without compromising the research question?

3. *Modify the dictionary.* If existing definitions do not adequately describe the scenario, and the scenario is deemed to be a critical aspect of the event as per the coding protocol, then a new option and operational definition must be added to the variable(s) used to describe those aspects of the event. If the research question does not require a unique description for the scenario, then, at minimum, the data dictionary needs to be modified so that an existing operational definition includes the scenario in question.

4. *Publish and distribute the updates.* Whenever the data dictionary is modified, all reductionists and researchers working with that dataset must be notified of the update as soon as possible. The most up-to-date data dictionary should be used at all times, and reductionists need to know that they are working with the most current definitions. Active notifications and reductionist compliance is strongly recommended. VTTI requires all reductionists to read and sign update documents after they have been afforded an opportunity to ask questions and clarify any uncertainties.

Phase 2: Coder Training

The tested and revised protocol enters the Coder Training phase. On the right side of the Phase 2 loop in Figure 1, the Senior Data Coder and the Data Coding Manager work together to train the first cohort of data Coders, ideally no more than 3-4 trainees at a time to keep the initial quality control manageable. The protocol is reviewed in detail with the Coders, and both paper and electronic copies are made available for reference. Once the formal training session has been completed, coders begin coding under the supervision of a Senior Coder. This initial coding should be short-term (no more than one full reduction shift), and then stopped for an accuracy assessment. Ideally, 100% of each Coders' work will be reviewed by a Senior Coder or Data Coder Manager. Corrections are made and detailed feedback is provided to Coders to review. These comments are reviewed with the Coders, and Coders are re-trained if necessary.

If Coders are able to meet reliability standards in this initial review (e.g., 90% accuracy, though this level may vary with the complexity and level of subjectivity present in the reduction), then a random sample (e.g., 10-20%) of all remaining work is checked (see Phase 3).

If the initial review was unsatisfactory, then another day of work is completed after retraining, and the 100% review is repeated. At this point, there may be times when all trainees are struggling with certain aspect or variables in the coding. This is usually a sign that either more in-depth training on that variable is required, or the protocol needs to be modified or clarified to increase reliability.

On the left side of the Phase 2 loop, the Senior Coder develops the first inter-rater test by selecting a sample of events that represent the range and frequency of conditions expected to be present in the data set. Depending on the complexity and length of the protocol, the test may include 10-30 events. After meeting the initial reliability standards during the 100% review, Coders are scheduled to take this test. By having all Coders complete the same set of events, the ability to consistently code variables within the group can be measured (inter-rater reliability). If scores on this test are satisfactory (e.g., 90% or greater), then Coders may move into Phase 3. If scores are unsatisfactory, retraining or additional protocol revisions may be necessary.

Finally, once the first cohort of 3-4 Coders completes the training loop and moves into Phase 3, additional groups of 3-4 Coders can enter the training loop. Depending on the protocol complexity and Coder experience, the training period may require approximately 2-5 days.

Phase 3: Data Coding

Three tools for ongoing quality control are used during the data coding phase: Spot Checks, Inter-rater tests, and Intra-rater tests.

A "spot check" consists of a second person reviewing previously completed data coding. Corrections can be made and comments provided to the original coder to confirm and/or improve accuracy. Spot checks are a very useful tool for assessing coders' understanding of protocols for a variety of potential circumstances, revealing potential ambiguities that may need to be addressed in the protocol or data dictionary, and monitoring overall data quality. Spot checks can be performed by the Data Coding Manager, Senior Data Coder, or experienced Data Coders (and usually a combination of all three). However, discrepancies between the original reduction and the spot check should generally be reviewed by a manager (Data Coding Manager or Senior Coder) as a third reviewer before changes are made. Feedback on all spot checks should be provided back to the original Coder immediately, and Coders then should review all comments and be encouraged to revisit the events and ask questions so that mistakes can be avoided in the future. If consistent or increasing errors are found during later spot checks, a larger sample of coding work for that individual should be checked and corrected until the issue is resolved.

Spot checks should begin very shortly after entering Phase 3. Generally, each person involved in spot checking will spend a certain amount of time each day/week performing spot checks (e.g., 1 hour per day, or more if necessary to complete the desired sample), and the rest of his/her time completing new coding. Spot checks should either be performed blindly so that the reviewer does not know who completed the original coding (in which case the reviewer may also review his/her own work) or openly so that each original coder can receive individualized feedback and errors can be easily contained and systematically corrected (in which case the reviewer should only review the work of others).

Inter-rater and Intra-rater tests are used periodically during a coding project to ensure consistency both between Coders at a given point in time (Inter) and within individual Coders over time (Intra). For the long-duration coding efforts such as the 100 Car Study, these tests were conducted once every month. This test is developed by the Senior Coder to include a sample of events that represents the range and frequency of conditions present in the data set. For the 100 Car, the Intra-rater test was conducted by repeating three of the events from the first test on each periodic test. The goal in this test was for each coder to record these events in the same way he/she coded it during the first test. If scores on this test were unsatisfactory, retraining occurred. Spot Checks and periodic Inter- and Intra-rater tests were continued until all events had been coded.

Phase 4: Data Delivery

In Phase 4, the data coding team works to prepare the data set for delivery back to the Researcher so that statistical analysis can begin. First, any remaining spots checks are completed and any remaining discrepancies between original coder and reviewer are resolved. Then, based on the spot check review and a pragmatic review of the protocol, all known errors or potential inconsistencies are reviewed. This is the Data Verification step. Examples of these issues might include a known confusion between the coding of a particular location as Interstate or Open Country in the Locality variable or a potential inconsistency between two variables that are present as a cross-check to assess internal consistency (e.g., if the subject is using a cell phone during an event, it should be marked as “Cell Phone” in the Distractions variable and “Distracted or Inattentive” in the Driver Behavior variable). With all of the QC measures taken during the first three phases, this step should be minimal, but is still critical. The Data Coding Manager and Senior Coders identify these issues and then work with the coders to resolve them. Data Verification is not complete until the data are internally consistent with the data dictionary. As a final review before data delivery, pulling all the data out into a spreadsheet and checking the relationship between events and variables is an additional recommended quality assurance step. Any questionable events are then flagged for a final review.

QA/QC Data Coding Results from the 100 Car Study

To determine the success of these techniques, an inter- and intra-rater reliability test was conducted during each month of data coding for the 100 Car Study. Three reliability tests were developed (each containing 20 events) for which the coders were required to make validity judgments (the most difficult variable to assess). Three of the 20 events were also completely re-reduced. Three of the test events on Test 1 were repeated on Test 2 and three other events were duplicated between Tests 2 and 3 to obtain a measure of intra-rater reliability.

Identifying Crash/Near-Crash Events

Using the expert reductionist's evaluations of each epoch as a "gold" standard, the proportion of agreement between the expert and each rater was calculated for each test. The measures for each rater for each testing period, along with a composite measure, can be found in Table 1.

<Insert Table S1 Here>

The Kappa statistic was also used to calculate inter-rater reliability. Although there is controversy surrounding the usefulness of the Kappa statistic, it is viewed by many researchers as the standard for rater assessment (e.g., Cicchetti and Feinstein, 1990). The Kappa coefficient (K = 0.65, p <0.0001) indicated that the association between raters is significant. While the coefficient value is somewhat low, given the highly subjective nature of the task, the number of raters involved, and the conservative nature of this statistic, the Kappa calculation errs on the low side.

A tetrachoric correlation coefficient is a statistical calculation of inter-rater reliability based on the assumption that the latent trait underlying the rating scale is continuous and normally distributed. Based on this assumption, the tetrachoric correlation coefficient can be interpreted in the same manner as a correlation coefficient calculated on a continuous scale. The average of the pair-wise correlation coefficients for the inter-rater analysis was 0.86 for the 100 Car Study. The coefficients for the intra-rater analysis were extremely high with nine raters achieving a correlation of 1.0 among the three reliability tests and five raters achieving a correlation of 0.99.

Given these three methods of calculating inter-rater reliability, it appears that the data reduction training coupled with spot-checking and weekly meetings proved to be an effective method for achieving high inter- and intra-rater reliability.

Intra-rater and inter-rater reliability were not conducted for the crash/near-crash events because these were coded by only two highly trained coders where both coders reviewed and coded all CNC events. 100% agreement was required for all CNC events. If there were discrepancies between these two coders, a senior researcher reviewed the event and made the final decision.

Control Road Segment Database for 100 Car Study

Three inter-rater reliability tests were also conducted for the baseline events. All trained data coders were given a random sample of 25 baseline epochs to view and record the secondary tasks, driving-related inattention behaviors, and fatigue. The coders' responses were then compared to an expert coder's responses. The results indicated an average of 88 percent accuracy among all the data coders. Given that neither the Kappa coefficient nor the tetrachoric correlation coefficient provided additional information, these tests were not conducted on the baseline inter-rater reliability test. Intra-rater scores for all coders averaged 95 percent accuracy.

QA/QC Data Coding Results from the Naturalistic Teenage Driving Study

Crash/Near Crash Coding

The crashes and near-crashes were coded by two highly trained coders. Both coders reviewed and recorded variables for all of the crashes and near-crashes. 100% agreement was required and any discrepancies were resolved by a senior researcher. Given that these events were reduced to such a high degree of accuracy, neither intra- nor inter-rater statistics were calculated for these events.

Control Road Segments

A single inter-rater reliability test was conducted for the control road segment coding because this coding was completed by eight coders in less than one month. The reliability test contained 15 baseline epochs for which the coders were required to make judgments regarding whether the driver was engaging in a distraction task, exhibiting signs of fatigue, or exhibiting signs of impairment. The results of this test are presented in Table 2. The average expert rater agreement was 93.3%, which suggests that the QA/QC process was successful for this study as well.

<Insert Table S2 Here>

Summary

The subjective coding of naturalistic driving video data is both one of the most powerful and rich data sources on driving safety but also one with high potential for error. It is because of the importance of these data that significant efforts were made to ensure high quality and efficiency in coding techniques. We believe that through these efforts, high quality subjective data have been gleaned for the studies reported here. We continue to develop and work on these methods to achieve highly accurate subjective data coding for these naturalistic studies but also future studies as well.

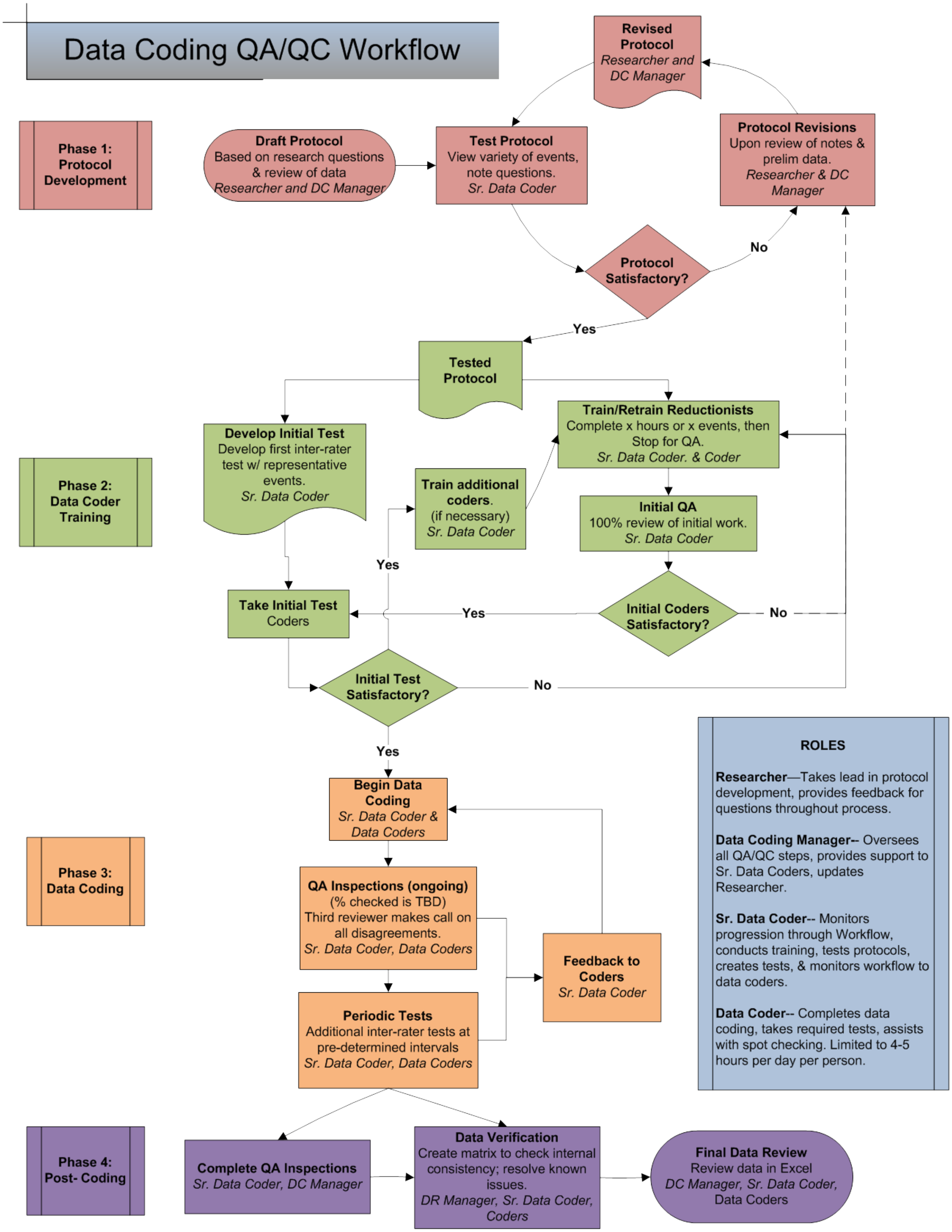
Data Coding QA/QC Workflow
Phase 1: Protocol Development
Draft Protocol
Based on research questions & review of data
Researcher and DC Manager
Test Protocol
View variety of events, note questions.
Sr. Data Coder
Revised Protocol
Researcher and DC Manager
Protocol Revisions
Upon review of notes & prelim data.
Researcher & DC Manager
Protocol Satisfactory?
No
Yes
Tested Protocol
Phase 2: Data Coder Training
Develop Initial Test
Develop first inter-rater test w/ representative events.
Sr. Data Coder
Train additional coders.
(if necessary)
Sr. Data Coder
Train/Retrain Reductionists
Complete x hours or x events, then Stop for QA.
Sr. Data Coder. & Coder
Initial QA
100% review of initial work.
Sr. Data Coder
Yes
Take Initial Test
Coders
Yes
Initial Coders Satisfactory?
No
Initial Test Satisfactory?
No
Yes
Phase 3: Data Coding
Begin Data Coding
Sr. Data Coder & Data Coders
QA Inspections (ongoing)
(% checked is TBD)
Third reviewer makes call on all disagreements.
Sr. Data Coder, Data Coders
Periodic Tests
Additional inter-rater tests at pre-determined intervals
Sr. Data Coder, Data Coders
Feedback to Coders
Sr. Data Coder
ROLES
Researcher—Takes lead in protocol development, provides feedback for questions throughout process.
Data Coding Manager-- Oversees all QA/QC steps, provides support to Sr. Data Coders, updates Researcher.
Sr. Data Coder-- Monitors progression through Workflow, conducts training, tests protocols, creates tests, & monitors workflow to data coders.
Data Coder-- Completes data coding, takes required tests, assists with spot checking. Limited to 4-5 hours per day per person.
Phase 4: Post- Coding
Complete QA Inspections
Sr. Data Coder, DC Manager
Data Verification
Create matrix to check internal consistency; resolve known issues.
DR Manager, Sr. Data Coder, Coders
Final Data Review
Review data in Excel
DC Manager, Sr. Data Coder, Data Coders

Figure S1. Data Coding QA/QC process and roles.

Table S1. Percentage agreement with expert reductionists.

Rater	Test 1 Percent	Test 2 Percent	Test 3 Percent
1	78.3	87.5	91.3
2	65.2	70.8	78.3
3	100	91.7	95.7
4	100	91.7	87.0
5	100	83.3	87.0
6	95.7	87.5	91.3
7	91.3	87.5	91.3
8	91.3	91.7	91.3
9	95.7	70.8	91.3
10	95.7	91.7	87.0
11	95.7	87.5	100
12	78.3	87.5	87.0
13	87.0	83.3	96.0
14	78.3	83.3	91.3
	Average (across all tests)	**88.4**	

Table S2. Results from the Inter-rater test conducted for NTD Study.

Rater	Test 1 Percent
1	100
2	93.3
3	100
4	93.3
5	86.7
6	86.7
7	100
8	86.7
Average (across all tests)	**93.3**

References

Cicchetti, D.V., and Feinstein, A.R. (1990). High agreement but low Kappa II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 45, 551-558.*

Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., Perez, M. A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z. R., Jermeland, J., and Knipling, R.R. (2006). *The 100-Car Naturalistic Driving Study: Phase II – Results of the 100-Car Field Experiment.* (Interim Project Report for DTNH22-00-C-07007, Task Order 6; Report No. DOT HS 810 593). Washington, D.C.: National Highway Traffic Safety Administration.

Lee, S. E., Simons-Morton, B. G., Klauer, S. G., Ouimet, M. C., and Dingus, T. A. (2011). Naturalistic Assessment of Novice Teenage Crash Experience. *Accident Analysis and Prevention*, *43*, 1472-1479.

Appendix 2: Table of Frequency Counts for Secondary Task Engagement for Novice and Experienced Drivers

Table 1. Frequency of engagement in secondary tasks in CNC and control segments

Secondary Task	**NTDS**		**100 Car Study**	
	Percent of CNC with secondary task	Percent of Control Road Segments	**Percent of CNC with secondary task**	Percent of Control Road Segments
Talking / Listening on cell phone	3.0%	4.56%	6.2%	6.1%
Dialing cell phone	3.0%	0.32%	2.7%	0.79%
Texting/ Email on cell phone	4.19%	1.35%	n/a	n/a
Reaching for cell phone	3.59%	0.4%	.04%	.19%
Reaching for Object (other than phone)	5.98%	0.82%	1.2%	1.5%
Eating	4.19%	1.4%	3.1%	1.8%
Drinking	1.20%	0.89%	.08%	1.5%
Other Vehicle Operations	2.40%	1.03%	.03%	.06%
Using Radio/HVA	7.19%	5.57%	2.5%	4.1%

C				
Roadside Object	4.79%	1.34%	2.3%	3.5%